Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature

Sándor Baran[1] and Annette Möller[2]
[1]Faculty of Informatics, University of Debrecen
Kassai út 26, H-4028 Debrecen, Hungary
[2] Department of Animal Sciences, University of Göttingen
Carl-Sprengel-Weg 1, D-37075 Göttingen, Germany

Abstract

Forecast ensembles are typically employed to account for prediction uncertainties in numerical weather prediction models. However, ensembles often exhibit biases and dispersion errors, thus they require statistical post-processing to improve their predictive performance. Two popular univariate post-processing models are the Bayesian model averaging (BMA) and the ensemble model output statistics (EMOS).

In the last few years increased interest has emerged in developing multivariate post-processing models, incorporating dependencies between weather quantities, such as for example a bivariate distribution for wind vectors or even a more general setting allowing to combine any types of weather variables.

In line with a recently proposed approach to model temperature and wind speed jointly by a bivariate BMA model, this paper introduces a bivariate EMOS model for these weather quantities based on a truncated normal distribution.

The bivariate EMOS model is applied to temperature and wind speed forecasts of the eight-member University of Washington mesoscale ensemble and of the eleven-member ALADIN-HUNEPS ensemble of the Hungarian Meteorological Service and its predictive performance is compared to the performance of the bivariate BMA model and a multivariate Gaussian copula approach, post-processing the margins with univariate EMOS. While the predictive skills of the compared methods are similar, the bivariate EMOS model requires considerably lower computation times than the bivariate BMA method.

Key words: Ensemble model output statistics, Gaussian copula, energy score, ensemble calibration, Euclidean error, truncated normal distribution.

1 Introduction

Accurate and reliable prediction of future states of the atmosphere is the most important objective of weather prediction. These forecasts are issued on the basis of observational data and numerical weather prediction (NWP) models, which are capable to simulate the atmospheric motions taking into account the physical governing laws of the atmosphere and the connected spheres (typically sea or land surface). The NWP models consist of sets of partial differential equations which have only numerical solutions and strongly depend on initial conditions. In order to reduce the uncertainties caused by the possibly unreliable initial conditions and the numerical weather prediction process itself, one can run the models with various initial conditions resulting in a forecast ensemble (Leith, 1974). Using a forecast ensemble not only the classical point forecasts (ensemble median or mean) can be obtained, but also an estimate of the distribution of the future weather variable, which allows probabilistic forecasting (Gneiting and Raftery, 2005). The first operational implementation of the ensemble prediction method dates back to the nineties (Buizza *et al.*, 1993; Toth and Kalnay, 1997) and in the last twenty years it became a widely used technique in the meteorological community. Recently all major national meteorological services operate their own ensemble prediction systems (EPSs), see, e.g., the PEARP[1] EPS of Météo France (Descamps *et al.*, 2014) or the COSMO-DE[2] EPS of the German Meteorological Service (DWD; Bouallègue *et al.*, 2013), whereas the most well-known organization issuing ensemble forecasts is the European Centre for Medium-Range Weather Forecasts (ECMWF Directorate, 2012). However, as it has been observed with several operational EPSs (see, e.g., Buizza *et al.*, 2005), the forecast ensemble is usually underdispersive and consequently badly calibrated. One possible improvement area of the ensemble forecasts is the statistical post-processing of the ensemble in order to transform the original ensemble member-based probability density function (PDF) into a more reliable and realistic one.

From the various post-processing techniques (for an overview see, e.g., Gneiting, 2014; Williams *et al.*, 2014) probably the most popular approaches are the Bayesian model averaging (BMA; Raftery *et al.*, 2005) and the ensemble model output statistics (EMOS) or non-homogeneous regression (Gneiting *et al.*, 2005). These methods are partially implemented in the `ensembleBMA` and `ensembleMOS` packages of `R` (Fraley *et al.*, 2011) and both approaches provide estimates of the distributions of the predictable weather quantities.

In the case of the BMA the predictive probability density function (PDF) of a future weather quantity is a weighted mixture of individual PDFs corresponding to the members of the ensemble, where the weights express the relative performance of the ensemble members during a given training period. The BMA models of various weather quantities differ only in the PDFs of the mixture components. For temperature and sea level pressure a normal distribution (Raftery *et al.*, 2005), for wind speed a gamma (Sloughter *et al.*, 2010) or a

[1] PEARP: Prévision d'Ensemble ARPege

[2] COSMO: Consortium for Small scale Modeling

truncated normal distribution (Baran, 2014), whereas for surface wind direction a von Mises distribution (Bao *et al.*, 2010) is suggested.

The EMOS predictive PDF uses a single parametric distribution with parameters depending on the ensemble members. EMOS models have already been developed for calibrating ensemble forecasts of temperature and sea level pressure (Gneiting *et al.*, 2005), wind speed (Thorarinsdottir and Gneiting, 2010; Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2015) and precipitation (Scheuerer, 2014).

Besides the calibration of univariate weather quantities recently an increasing interest has appeared in modeling correlations between the different weather variables. In the special case of wind vectors Pinson (2012) suggested an adaptive calibration technique, whereas Schuhen *et al.* (2012) and Sloughter *et al.* (2013) introduced bivariate EMOS and BMA models, respectively. Further, Möller *et al.* (2013) developed a general approach where after univariate calibration of the weather variables the component predictive PDFs are joined into a multivariate predictive density with the help of a Gaussian copula. Another idea appears in the ensemble copula coupling (ECC) method Schefzik *et al.* (2013) where after univariate calibration the rank order information in the raw ensemble is used to restore correlations. Finally, Baran and Möller (2015) developed a BMA model for joint post-processing of ensemble forecasts of wind speed and temperature.

In the present paper we introduce an EMOS model for joint calibration of wind speed and temperature which is based on a truncated normal distribution with cut-off at zero in its first (wind) coordinate. The method is tested on the ensemble forecasts of wind speed and temperature of the eight-member University of Washington Mesoscale Ensemble (UWME; Eckel and Mass, 2005) and of the Limited Area Model EPS of the Hungarian Meteorological Service (HMS) called ALADIN-HUNEPS[3] (Horányi *et al.*, 2011). The performance of the EMOS model is compared to the forecasting skills of the previously investigated BMA method of Baran and Möller (2015) and to the Gaussian copula approach of Möller *et al.* (2013), where the margins of the multivariate predictive distribution are estimated by EMOS.

2 Data

2.1 University of Washington mesoscale ensemble

The eight-member University of Washington mesoscale ensemble covers the Pacific Northwest region of western North America providing forecasts on a 12 km grid. The ensemble members are obtained from different runs of the fifth generation Pennsylvania State University–National Center for Atmospheric Research mesoscale model (PSU-NCAR MM5) with initial conditions from different sources (Grell *et al.*, 1995). Our data base (identical to the one used in Möller *et al.* (2013); Baran and Möller (2015)) contains ensembles of 48-

[3]ALADIN: Aire Limitée Adaptation dynamique Development International

hour forecasts and corresponding validation observations of 10 meter maximum wind speed (maximum of the hourly instantaneous wind speeds over the previous twelve hours, given in m/s, see, e.g., Sloughter *et al.* (2010)) and 2 meter minimum temperature (given in K) for 152 stations in the Automated Surface Observing Network (National Weather Service, 1998) in the US states of Washington, Oregon, Idaho, California and Nevada for calendar years 2007 and 2008. The forecasts are initialized at 0 UTC (5 pm local time when daylight saving time (DST) is in use and 4 pm otherwise) and the generation of the ensemble implies that its members are not exchangeable. In the present study we investigate only forecasts for calendar year 2008 with additional data from 2007 used for parameter estimation. After removing days and locations with missing data, 90 stations remained where the number of days for which forecasts and validating observations are available varies between 141 and 290.

2.2 ALADIN-HUNEPS ensemble

The ALADIN-HUNEPS system of the HMS covers a large part of Continental Europe with a horizontal resolution of 8 km and it is obtained by dynamical downscaling (by the ALADIN limited area model) of the global ARPEGE[4] based PEARP system of Météo France (Horányi *et al.*, 2006; Descamps *et al.*, 2014). The ensemble consists of 11 members, 10 initialized from perturbed initial conditions and one control member from the unperturbed analysis, implying that the ensemble contains groups of exchangeable forecasts. The data base contains 11 member ensembles of 42-hour forecasts for 10 meter instantaneous wind speed (given in m/s) and 2 meter temperature (given in K) for 10 major cities in Hungary (Miskolc, Szombathely, Győr, Budapest, Debrecen, Nyíregyháza, Nagykanizsa, Pécs, Kecskemét, Szeged) produced by the ALADIN-HUNEPS system of the HMS, together with the corresponding validating observations for the one-year period between April 1, 2012 and March 31, 2013 and for the period from October 1, 2010 to March 25, 2011. The forecasts are initialized at 18 UTC (8 pm local time when DST operates and 7 pm otherwise). The data sets are fairly complete since there are only six and three days, respectively, when no forecasts are available and these days have been excluded from the analysis.

3 Ensemble Model Output Statistics

As mentioned in the Introduction, the EMOS predictive PDF of a weather quantity (vector) X is a single parametric density function where the parameters depend on the ensemble members. For temperature and pressure a normal distribution can be fit reasonably well (Gneiting *et al.*, 2005), while for wind vectors a bivariate normal distribution can be applied (Schuhen *et al.*, 2012). However, for modeling non-negative quantities such as wind

[4]ARPEGE: Action de Recherche Petite Echelle Grande Echelle

speed, a skewed distribution is required. Thorarinsdottir and Gneiting (2010) introduced an EMOS model based on truncated normal distribution with cut-off at zero, but EMOS models utilizing a generalized extreme value distribution (Lerch and Thorarinsdottir, 2013) and a log-normal distribution (Baran and Lerch, 2015) have also been tested. The EMOS models of Gneiting *et al.* (2005) and Thorarinsdottir and Gneiting (2010) suggest the idea of joint modeling wind speed and temperature using a bivariate normal distribution with first (wind) coordinate truncated from below at zero. This particular distribution has already been applied in the bivariate BMA model of Baran and Möller (2015).

Once the predictive density is given, its mean or median can be taken as a point forecast for X. In one dimension the definition of the latter is obvious, whereas for a d-dimensional cumulative distribution function (CDF) F a multivariate median is a vector minimizing the function

$$\phi(\boldsymbol{\alpha}) := \int_{\mathbb{R}^d} \|\boldsymbol{\alpha} - \boldsymbol{x}\| F(\mathrm{d}\boldsymbol{x}),$$

where $\|\cdot\|$ denotes the Euclidean norm. If F is not concentrated on a line in $\mathbb{R}^d$ then the median is unique (Milasevic and Ducharme, 1987).

Denote by $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_M$ the ensemble of distinguishable forecast vectors of wind speed and temperature for a given location and time. This means that each ensemble member can be identified and tracked, which holds for example for the UWME (see Section 2.1). However, most of the currently used ensemble prediction systems provide ensembles where at least some members are statistically indistinguishable. Such ensemble systems are simulating uncertainties by perturbing the initial conditions, and they usually have a control member (the one without any perturbation), whereas the remaining ensemble members form one or two exchangeable groups. This is the case, e.g., for the 51 member ECMWF ensemble (Leutbecher and Palmer, 2008) or for the ALADIN-HUNEPS ensemble described in Section 2.2.

In what follows, if we have M ensemble members divided into m exchangeable groups, where the kth group contains $M_k \geq 1$ ensemble members ($\sum_{k=1}^{m} M_k = M$), notation $\boldsymbol{f}_{k,\ell}$ will be used for the ℓth member of the kth group.

3.1 Bivariate truncated normal model

Denote by $\mathcal{N}_2^0(\boldsymbol{\mu}, \Sigma)$ the bivariate normal distribution with location vector $\boldsymbol{\mu}$, scale matrix Σ, and first coordinate truncated from below at zero. Let

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_W \\ \mu_T \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \sigma_W^2 & \sigma_{WT} \\ \sigma_{WT} & \sigma_T^2 \end{bmatrix}.$$

If Σ is regular, then the PDF of this distribution is

$$g(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) := \frac{\big(\det(\Sigma)\big)^{-1/2}}{2\pi\Phi\big(\mu_W/\sigma_W\big)} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big)\mathbb{I}_{\{x_W \geq 0\}}, \quad \boldsymbol{x} = \begin{bmatrix} x_W \\ x_T \end{bmatrix} \in \mathbb{R}^2, \quad (3.1)$$

where Φ denotes the CDF of the standard normal distribution and by $\mathbb{I}_H$ we denote the indicator function of a set H. Short calculation shows (see, e.g., Rosenbaum, 1961), that the mean vector $\boldsymbol{\kappa}$ and covariance matrix Ξ of $\mathcal{N}_2^0(\boldsymbol{\mu}, \Sigma)$ are

$$\boldsymbol{\kappa} = \boldsymbol{\mu} + \frac{\varphi\big(\mu_W/\sigma_W\big)}{\Phi\big(\mu_W/\sigma_W\big)} \begin{bmatrix} \sigma_W \\ \sigma_{WT}/\sigma_W \end{bmatrix} \qquad \text{and}$$

$$\Xi = \Sigma - \left(\frac{\mu_W}{\sigma_W} \frac{\varphi\big(\mu_W/\sigma_W\big)}{\Phi\big(\mu_W/\sigma_W\big)} + \left(\frac{\varphi\big(\mu_W/\sigma_W\big)}{\Phi\big(\mu_W/\sigma_W\big)} \right)^2 \right) \begin{bmatrix} \sigma_W^2 & \sigma_{WT} \\ \sigma_{WT} & \sigma_{WT}^2/\sigma_W^2 \end{bmatrix},$$

respectively, where φ denotes the PDF of the standard normal distribution.

The proposed EMOS predictive distribution of wind speed and temperature is

$$\mathcal{N}_2^0\big(A + B_1\boldsymbol{f}_1 + \cdots + B_M\boldsymbol{f}_M, C + DSD^\top\big) \tag{3.2}$$

with

$$S := \frac{1}{M-1} \sum_{k=1}^{M} \big(\boldsymbol{f}_k - \overline{\boldsymbol{f}}\big)\big(\boldsymbol{f}_k - \overline{\boldsymbol{f}}\big)^\top,$$

where $\overline{\boldsymbol{f}}$ denotes the ensemble mean vector. Parameter vector $A \in \mathbb{R}^2$ and two-by-two real parameter matrices $B_1, \ldots, B_M$ and C, D of model (3.2), where C is assumed to be symmetric and non-negative definite, can be estimated from the training data consisting of ensemble members and verifying observations from the preceding n days, by optimizing with respect to the mean logarithmic score, i.e., the negative logarithm of the predictive PDF evaluated at the verifying observation (Gneiting *et al.*, 2008). We remark that under the assumption of independence in space and time, this approach is equivalent to the maximum likelihood method. Obviously, the forecast errors are usually not independent, however, since one is estimating the conditional distribution of a single weather quantity vector with respect to the corresponding forecasts, the parameter estimates are not really sensitive to this assumption (see, e.g., Raftery *et al.*, 2005).

If the ensemble can be divided into groups of exchangeable members, ensemble members within a given group will get the same coefficient matrix of the location parameter (Fraley *et al.*, 2010; Gneiting, 2014) resulting in a predictive distribution of the form

$$\mathcal{N}_2^0\left(A + B_1 \sum_{\ell_1=1}^{M_1} \boldsymbol{f}_{1,\ell_1} + \cdots + B_m \sum_{\ell_m=1}^{M_m} \boldsymbol{f}_{m,\ell_m}, C + DSD^\top \right), \tag{3.3}$$

where again, S denotes the empirical covariance matrix of the ensemble.

3.2 Verification scores

To investigate the predictive skills of the probabilistic and point forecasts we apply the multivariate scores proposed by Gneiting *et al.* (2008).

The first step is to check the calibration of probabilistic forecasts, which notion means a statistical consistency between the predictive distributions and the observations (see, e.g., Thorarinsdottir and Gneiting, 2010). For one-dimensional ensemble forecasts a frequently used tool for this purpose is the verification rank histogram, i.e., the histogram of ranks of validating observations with respect to the ensemble forecasts (see, e.g., Wilks, 2011, Section 8.7.2). The closer the distribution of the ranks to the uniform distribution on $\{1, 2, \ldots, M+1\}$, the better the calibration. The deviation from uniformity can be quantified by the reliability index Δ defined as

$$\Delta := \sum_{r=1}^{M+1} \Big| \rho_r - \frac{1}{M+1} \Big|,$$

where ρ_j is the relative frequency of rank r (Delle Monache *et al.*, 2006). In the multivariate case the proper definition of ranks is not obvious. Similar to Baran and Möller (2015), in the present work we use the multivariate ordering proposed by Gneiting *et al.* (2008). For a probabilistic forecast one can calculate the reliability index from a preferably large number of ensembles (we use 100) sampled from the predictive PDF and the corresponding verifying observations.

For evaluating multivariate density forecasts the most popular scoring rules are the logarithmic score and the energy score (ES), introduced by Gneiting and Raftery (2007). Both the logarithmic and the energy score are proper scoring rules which are negatively oriented, that is, the smaller the better, and the latter is a direct multivariate extension of the continuous ranked probability score (CRPS). Given a predictive CDF F on $\mathbb{R}^d$ and a d-dimensional observation $\boldsymbol{x}$, the energy score is defined as

$$\mathrm{ES}(F, \boldsymbol{x}) := \mathsf{E}\|\boldsymbol{X} - \boldsymbol{x}\| - \frac{1}{2}\mathsf{E}\|\boldsymbol{X} - \boldsymbol{X}'\|,$$

where $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors with CDF F. However, for the bivariate truncated normal distribution the energy score cannot be given in a closed form, so it is replaced by a Monte Carlo approximation

$$\widehat{\mathrm{ES}}(F, \boldsymbol{x}) := \frac{1}{n}\sum_{j=1}^{n} \|\boldsymbol{X}_j - \boldsymbol{x}\| - \frac{1}{2(n-1)}\sum_{j=1}^{n-1} \|\boldsymbol{X}_j - \boldsymbol{X}_{j+1}\|, \tag{3.4}$$

where $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a (large, we use $n = 10000$) random sample from F (Gneiting *et al.*, 2008). Finally, if F is a CDF corresponding to a forecast ensemble $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_M$ then (3.4) reduces to

$$\mathrm{ES}(F, \boldsymbol{x}) = \frac{1}{M}\sum_{j=1}^{M} \|\boldsymbol{f}_j - \boldsymbol{x}\| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} \|\boldsymbol{f}_j - \boldsymbol{f}_k\|.$$

Besides the proper calibration, probabilistic forecasts should result in sharp predictive distributions. In the univariate case this usually means small standard deviations leading

to narrow central prediction intervals. For a d-dimensional quantity one can consider the determinant sharpness (DS) defined by

$$\mathrm{DS} := \big(\det(\Sigma)\big)^{1/(2d)},$$

where Σ is the covariance matrix of an ensemble or of a predictive PDF.

Finally, point forecasts (median and mean) can be evaluated using the mean Euclidean distance (EE) of forecasts from the corresponding validating observations. For multivariate forecasts the ensemble median can be obtained, e.g., using the Newton-type algorithm given in Dennis and Schnabel (1983) or the algorithm of Vardi and Zhang (2000). For a detailed comparison of different algorithms, see, e.g., Fritz *et al.* (2012). Given a predictive CDF, to determine the corresponding median the chosen algorithm might be applied on a preferably large sample from this distribution.

3.3 Parameter estimation

There are two possible approaches to the choice of training data for estimating the unknown parameters of the various EMOS models (Thorarinsdottir and Gneiting, 2010; Schuhen *et al.*, 2012). The regional EMOS technique uses ensemble forecasts and validating observations from a rolling training period for all available stations. In this way, one gets a universal set of parameters across the entire ensemble domain, which is then used at all observation sites. E.g., in case of the ALADIN-HUNEPS ensemble this means a single set of parameters for all ten cities. In contrast, local EMOS produces distinct parameter estimates for the different stations by using only the training data of the given station. These training sets contain only one observation per day, so local EMOS models require long training periods.

Now, e.g., in the bivariate model (3.3) the number of free parameters to be estimated is $4m + 10$, which means 14 parameters even in the simplest case of a single exchangeable ensemble group. Hence, for estimating the parameters of models (3.2) and (3.3) only the regional EMOS approach is applicable.

The mean logarithmic score is optimized numerically with the help of the `optim` function in `R`, using principally the Nelder-Mead (Nelder and Mead, 1965) algorithm. This method is slower but more robust than the popular Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Press *et al.*, 2007, Section 10.9), which in case of a small training set becomes unstable. Both optimization methods require initial values, and the starting values of the location parameters A and $B_1, \ldots, B_M$ are coefficients of the bivariate linear regression of the observations on the ensemble forecasts over the training period. Further, for the scale parameters C and D, the previous day's estimates can serve as initials values, however, according to our experience, fixed starting values provide slightly better results. Finally, to enforce the non-negative definiteness of the parameter matrix C one can set $C = \mathcal{C}\mathcal{C}^\top$ and perform the optimization with respect to $\mathcal{C}$.

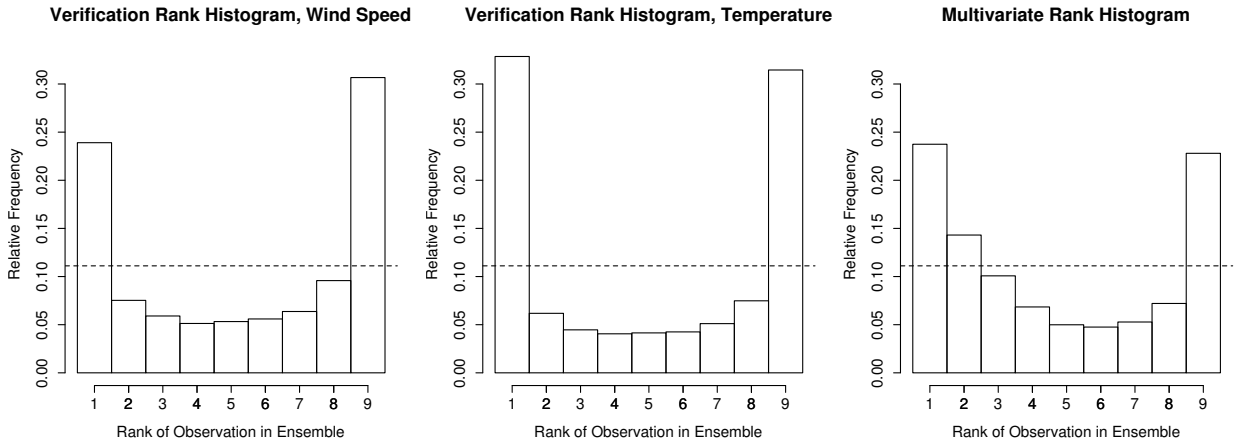


Figure 1: Verification rank histograms of the 8-member UMWE forecasts of maximum wind speed (*left*) and minimum temperature (*center*) and the multivariate rank histogram (*right*). Period: January 1, 2008 – December 31, 2008.

4 Results

As mentioned in the Introduction, the predictive performance of the bivariate EMOS model (see Section 3.1) is tested on the eight-member UWME and on the ALADIN-HUNEPS ensemble of the HMS. The goodness of fit of the predictive distributions is quantified with the multivariate scores given in Section 3.2, and the obtained results are compared to the fits of the independent EMOS models of wind speed (Thorarinsdottir and Gneiting, 2010) and temperature (Gneiting *et al.*, 2005), the Gaussian copula method proposed by Möller *et al.* (2013), but with marginal distributions estimated by EMOS models, and the bivariate BMA model of Baran and Möller (2015). We remark that the parameters of the independent univariate EMOS models are estimated by minimizing the mean CRPS of the training data. For fitting the marginal predictive distributions in the Gaussian copula approach, we employ the same univariate EMOS models for wind speed and temperature as in the independent approach. Therefore, their model parameters are estimated by the minimum CRPS method as well. If one has a closed expression for the CRPS, which is the case both for the normal and the truncated normal distribution, this method usually gives better results than optimization with respect to the logarithmic score.

4.1 University of Washington Mesoscale Ensemble

4.1.1 Raw ensemble

Several studies have verified that wind speed and temperature forecasts of the UWME are strongly underdispersive (see, e.g., Thorarinsdottir and Gneiting, 2010; Fraley *et al.*, 2010), and consequently uncalibrated. Obviously, the lack of calibration will remain valid if one

	Probabilistic forecasts			Median forecasts			Mean forecasts		
	ES	Δ	DS	EE	ϱ	ϱ_{err}	EE	ϱ	ϱ_{err}
EMOS	2.127	0.025	2.273	2.982	0.165	0.182	2.982	0.157	0.182
Indep. EMOS	2.118	0.059	2.206	2.966	0.164	0.176	2.966	0.155	0.178
Copula	2.088	0.021	2.169	2.967	0.162	0.178	2.967	0.156	0.179
BMA	2.110	0.015	2.250	2.973	0.154	0.182	2.972	0.155	0.183
Raw ensemble	2.562	0.550	0.773	3.087	0.017	0.187	3.072	0.007	0.189

Table 1: Mean energy score (ES), reliability index (Δ) and mean determinant sharpness (DS) of probabilistic forecasts, mean Euclidean error (EE) of point forecasts (median/mean), empirical correlation (ϱ) and empirical correlation of errors (ϱ_{err}) of wind speed and temperature components of point forecasts for the UWME. Empirical correlation of observations corresponding to the forecast cases: 0.125.

considers these ensemble forecasts together, as predictions of a bivariate weather quantity (Baran and Möller, 2015). The underdispersive character of the raw ensemble can nicely be observed in Figure 1 (identical to Figure 1 of Baran and Möller (2015)) displaying the univariate verification rank histograms of wind speed and temperature forecasts together with their joint multivariate rank histogram. The corresponding reliability indices Δ are 0.647, 0.842 and 0.550, respectively, and in many cases the raw ensemble either over-, or underestimates the verifying observation. Further, the need of bivariate modeling can be justified both by the positive correlation of 0.125 of the verifying observations of wind speed and temperature for calendar year 2008 taken along all dates and locations, and by the correlations of 0.187 and 0.189 of forecast errors of the ensemble median and mean, respectively.

4.1.2 Bivariate EMOS calibration

The first step of EMOS (and BMA) post-processing of ensemble forecast is the selection of the length of the rolling training period. In order to ensure comparability of the results with the findings of earlier studies, we apply the same 40 days training period length as in Möller *et al.* (2013) and Baran and Möller (2015). This training period length was a result of an exploratory data analysis on a subset of the data set. Similar to the previous studies, we produce EMOS predictive PDFs for the whole calendar year 2008, using also the data from the last two months of calendar year 2007. After removing dates with missing data this means 291 calendar days with a total of 24 302 forecast cases. As the eight ensemble members of the UWME are not exchangeable, for calibration we apply bivariate EMOS model (3.2) with $M = 8$.

In case of the copula method, the data from calendar year 2007 are applied for estimating the correlation between the two weather quantities, and the resulting correlation matrix is

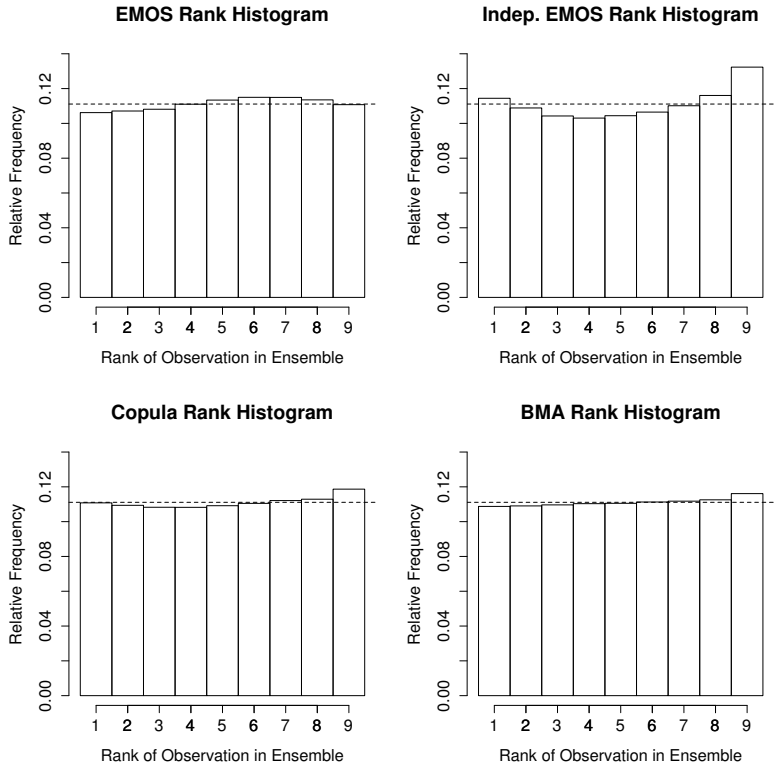


Figure 2: Multivariate rank histograms for EMOS (*upper left*), independent EMOS (*upper right*), Gaussian copula (*lower left*) and BMA (*lower right*) post-processed UWME forecasts of maximum wind speed and minimum temperature.

then carried forward into the analysis of the 2008 data. This is in accordance with the BMA based copula calibration of Baran and Möller (2015).

In Table 1 the verification scores calculated using the EMOS model (3.2), the independent EMOS models of wind speed and temperature, the copula model of Möller *et al.* (2013) with EMOS post-processed margins, the BMA model of Baran and Möller (2015) and the raw ensemble are given. Compared to the raw ensemble all post-processing techniques substantially improve the calibration of probabilistic forecasts which is quantified by the significant decrease of the mean energy score (ES) and reliability index (Δ) and can also be observed in Figure 2 showing the rank histograms of post-processed forecasts. These almost uniform histograms should be compared to the rank histograms of the raw ensemble plotted in Figure 1. The price to pay for the better calibration is the loss in sharpness (see the corresponding values of DS), however, this is a direct consequence of the small dispersion of the raw ensemble (see again Figure 1). Post-processing also results in slightly smaller mean Euclidean errors (EE) indicating more accurate median and mean forecasts. Further, the empirical correlations ϱ of the wind and temperature components of the post-processed point forecasts are much closer to the correlation of 0.125 of the verifying observations than the corresponding correlations of the ensemble median and mean. This latter is a weakness of

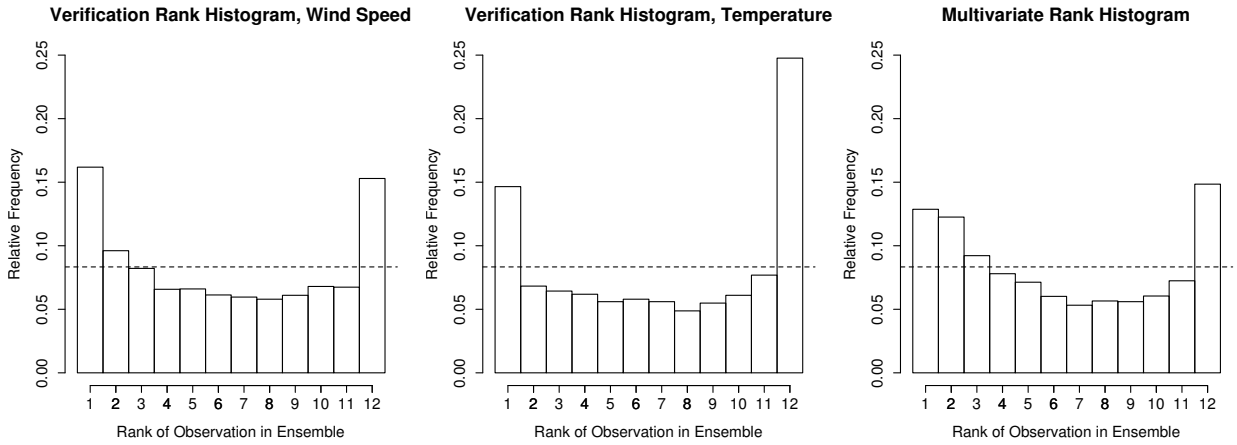


Figure 3: Verification rank histograms of the 11-member ALADIN-HUNEPS ensemble forecasts of wind speed (*left*) and temperature (*center*) and the multivariate rank histogram (*right*). Period: April 1, 2012 – March 31, 2013.

the raw ensemble, however, one should also remark that all error correlations ϱ_{err} (including the raw ensemble) are very similar to each other (around 0.180).

Comparing the different post-processing techniques one can observe that the main difference between the various approaches appears in the reliability index. The bivariate BMA model results in the smallest Δ value, followed by the copula and the bivariate EMOS methods, which is in line with shapes of the multivariate rank histograms plotted in Figure 2. Further, the large Δ value and the U-shaped rank histogram of the independent EMOS approach supports the idea of bivariate modeling. However, in the model choice one should also take into account that the copula method requires additional data for estimating the correlation matrix, whereas in the BMA and EMOS approaches the parameters are estimated using only the training data. Finally, in case of the latter two methods the computational costs (see Section 4.3) might also have an influence on the decision.

4.2 ALADIN-HUNEPS Ensemble

4.2.1 Raw ensemble

Wind speed and temperature forecasts of the ALADIN-HUNEPS EPS are better calibrated than those of the UWME, however, the rank histograms in Figure 3 still exhibit a strong underdispersive character. The bivariate reliability index equals 0.317, whereas the reliability indices of wind speed and temperature are 0.322 and 0.455, respectively. The need of bivariate post-processing is again supported by the forecast error correlations of 0.119 and 0.123 of the ensemble median and mean, respectively, however, in this case the verifying observations of wind speed and temperature show a very slight negative correlation of -0.029.

	Probabilistic forecasts			Median forecasts			Mean forecasts		
	ES	Δ	DS	EE	ϱ	ϱ_{err}	EE	ϱ	ϱ_{err}
EMOS	1.442	0.034	1.478	2.015	-0.041	0.132	2.016	-0.049	0.132
Indep. EMOS	1.436	0.051	1.456	2.002	-0.033	0.128	2.002	-0.044	0.127
Copula	1.384	0.075	1.557	2.000	-0.036	0.128	2.000	-0.039	0.127
BMA	1.434	0.031	1.539	2.004	-0.032	0.129	2.007	-0.041	0.129
Raw ensemble	1.623	0.327	0.935	2.102	-0.068	0.122	2.083	-0.060	0.124

Table 2: Mean energy score (ES), reliability index (Δ) and mean determinant sharpness (DS) of probabilistic forecasts, mean Euclidean error (EE) of point forecasts (median/mean), empirical correlation (ϱ) and empirical correlation of errors (ϱ_{err}) of wind speed and temperature components of point forecasts for the ALADIN-HUNEPS ensemble. Empirical correlation of observations corresponding to the forecast cases: -0.033.

This latter difference compared to the UWME, where this correlation equals 0.125, might be explained by the different types of wind and temperature quantities being examined (see Sections 2.1 and 2.2).

4.2.2 Bivariate EMOS calibration

Similar to the case of the UWME, to ensure the comparability of the results with the bivariate BMA post-processing of the same forecast data, we keep the 40-day training period of Baran and Möller (2015). This particular training period length was the outcome of a preliminary data analysis consisting of univariate BMA and EMOS calibration of wind speed and temperature forecasts. Hence, ensemble forecasts, validating observations and predictive distributions are available for the period from May 12, 2012 to March 31, 2013, which means 318 days and 3 180 forecast cases as 6 days with missing forecasts are excluded from the analysis.

Further, the way the ALADIN-HUNEPS ensemble is generated (see Section 2.2) induces a natural grouping of the ensemble members into two groups. The first group contains just the control member $\boldsymbol{f}_c$, whereas in the second are the 10 statistically indistinguishable ensemble members $\boldsymbol{f}_{p,1}, \ldots, \boldsymbol{f}_{p,10}$, initialized from randomly perturbed initial conditions. This results in the predictive PDF

$$\mathcal{N}_2^0\left(A + B_c\boldsymbol{f}_c + B_p\sum_{\ell=1}^{10}\boldsymbol{f}_{p,\ell}, C + DSD^\top\right), \tag{4.1}$$

which is a special case of model (3.3). One should remark here that in Baran *et al.* (2013) a different grouping is also suggested (and later investigated in Baran (2014); Baran *et al.*

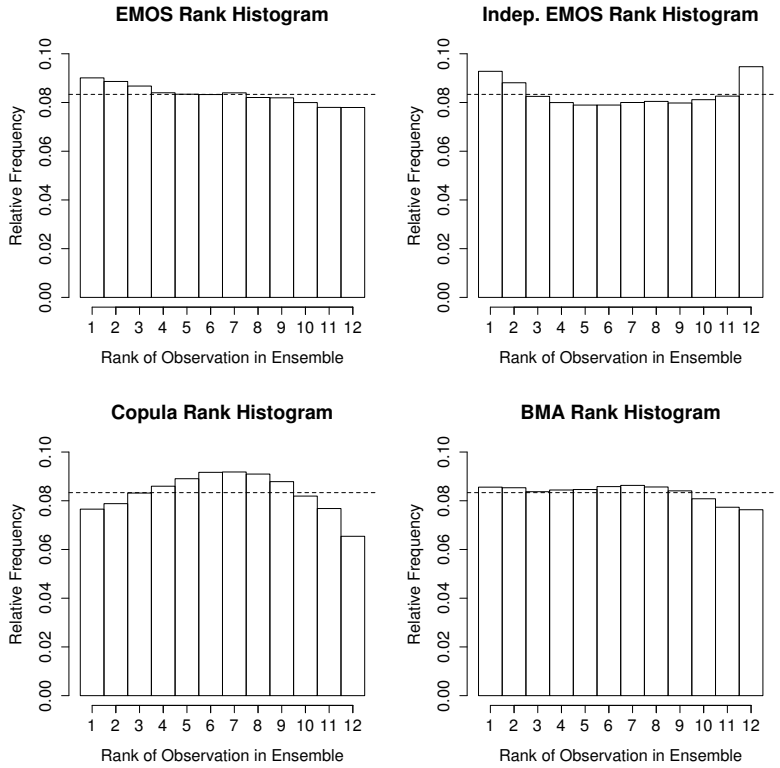


Figure 4: Multivariate rank histograms for EMOS (*upper left*), independent EMOS (*upper right*), Gaussian copula (*lower left*) and BMA (*lower right*) post-processed ALADIN-HUNEPS forecasts of instantaneous wind speed and temperature.

(2014) and Baran and Möller (2015), too), where the odd and even numbered exchangeable ensemble members form two separate groups. This idea is justified by the method their initial conditions are generated, since only five perturbations are calculated and then they are added to (odd numbered members) and subtracted from (even numbered members) the unperturbed initial conditions. However, since in the present study the results corresponding to the two- and three-group models are rather similar, only the two-group case is reported.

In line with the similar case study of Baran and Möller (2015), to estimate the correlation matrix of the Gaussian copula, additional data of the period from October 1, 2010 to March 25, 2011 are utilized, and the estimated correlation matrix is employed for combining the univariate EMOS marginals for 2012/2013 in the Gaussian copula.

The effects of statistical calibration of ensemble forecasts are quantified by the multivariate scores given in Table 2. Compared to the raw ensemble all four post-processing methods result in significantly lower energy scores and reliability indices (compare Figures 3 and 4) and higher DS values. Again, the loss in determinant sharpness is an effect of the underdispersive nature of the ensemble. However, here the increase in DS is around 60 %, whereas for the UWME the raw ensemble is almost three times sharper than the various predictive PDFs. This again indicates the better calibration of the ALADIN-HUNEPS ensemble

which is fully consistent with Figures 1 and 3 and the corresponding reliability indices given in Sections 4.1.1 and 4.2.1, respectively. Further, the ensemble median and mean vectors produce slightly larger Euclidean errors than the corresponding post-processed point forecasts. Moreover, the empirical correlations of the components of the ensemble median and mean are almost the double of the nominal correlation -0.033 of observations, whereas the correlations of wind speed and temperature components of the BMA and EMOS point forecasts are close to this value. Finally, both the ensemble median/mean and their calibrated counterparts exhibit almost the same forecast error correlations.

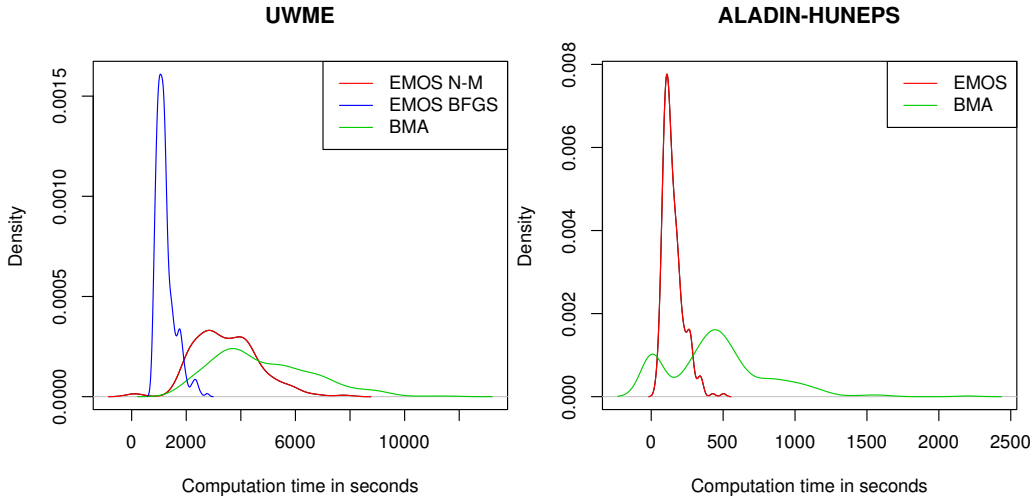
From the competing post-processing methods the Gaussian copula approach results in the lowest energy score and Euclidean errors, however, the differences compared to the corresponding scores of the BMA and EMOS models (especially in the EE values) are rather small. Reliability indices show far larger variability and the highest scores belong to the copula model and to the independent EMOS approach. The Δ values in Table 2 are in accordance with the rank histograms in Figure 4: the rank histogram of the copula method is strongly hump-shaped indicating over-dispersion (see, e.g., Gneiting *et al.*, 2008), whereas the histogram of the independent EMOS approach exhibits some under-dispersion. For the ALADIN-HUNEPS ensemble the bivariate BMA model has the best overall performance closely followed by the bivariate EMOS method, however, similar to the case of the UWME the computational costs might also effect the model choice.

4.3 Computational aspects

For all EMOS methods which have been developed so far the most time-consuming and problematic part of ensemble post-processing is the numerical optimization used in parameter estimation. In case of bivariate EMOS calibration of the ALADIN-HUNEPS ensemble only the robust Nelder-Mead algorithm occurs to be reliable, as one has to estimate 18 free parameters with the help of 400 forecast cases of the training data. For the UWME the data/parameter ratio is much better, as 42 free parameters have to be estimated using on average 3354 forecast cases. For this data set the reported Nelder-Mead and the faster BFGS algorithm give almost the same results.

In case of the BMA calibration the bottleneck with respect to the computation costs is the EM algorithm applied for ML estimation of the parameters. The bivariate BMA model of Baran and Möller (2015) makes use of a modification of the truncated data EM algorithm for Gaussian mixture models (Lee and Scott, 2012) which operates with closed formulae and there is no need of numerical optimization. However, due to the large number of free parameters (UWME: 59; ALADIN-HUNEPS: 17) it requires quite a lot of iterations resulting in long computation times.

The Gaussian copula method starts with very fast univariate EMOS calibration, however, it utilizes an additional data set for estimating the correlation matrix of the Gaussian copula and additional post-processing of the univariate predictive PDFs. Hence, in terms of



(a) (b)

Figure 5: Densities of computation times for the bivariate BMA and EMOS models. a) UWME for the calendar year 2008; b) ALADIN-HUNEPS ensemble for the period May 12, 2012 – March 31, 2013.

computational efficiency this method is not comparable with the bivariate approaches and it is excluded from our analysis.

Figures 5a and 5b show the kernel density estimates of the distribution of computation times over the days in the verification period for bivariate BMA and EMOS models (implemented in `R`) for the UWME and ALADIN-HUNEPS ensemble, respectively, calculated on a portable computer under a 64 bit Fedora 20 operating system (Intel Quad Core i7-4700MQ CPU (2.40GHz $\times$ 4), 20 Gb RAM). We remark that in Figure 5a the density of computation times of the EMOS model with BFGS optimization is also plotted. The densities displayed in Figure 5 clearly show that in terms of computation time the EMOS model outperforms the BMA approach. The same conclusion can be derived from Table 3 where the median, mean and standard deviation of the computation times are given. However, one should also remark that these computation times are still too long for an operational use.

5 Conclusions

We introduce a new EMOS model for joint calibration of ensemble forecasts of wind speed and temperature providing a predictive PDF which follows a bivariate normal distribution truncated from below at zero in its first coordinate. The model is tested on wind speed and temperature forecasts of the eight-member University of Washington mesoscale ensemble and of the eleven-member ALADIN-HUNEPS ensemble of the Hungarian Meteorological Service. These ensemble prediction systems differ both in the weather quantities being forecasted and in the generation of the ensemble members.

Model	UWME			ALADIN-HUNEPS	
	EMOS		BMA	EMOS	BMA
	Nelder-Mead	BFGS		Nelder-Mead	
median	3349.443	1140.702	4419.288	131.609	436.873
mean	3475.681	1228.142	4801.247	150.008	459.560
std. dev.	1177.651	343.633	1823.083	69.678	345.187

Table 3: Median, mean and standard deviation of the computation times in seconds allocated to the parameter estimation for individual days in the verification period (UWME: calendar year 2008, 24 302 forecast cases; ALADIN-HUNEPS ensemble: May 12, 2012 – March 31, 2013, 3 180 forecast cases).

Using appropriate verification measures (energy score, reliability index and determinant sharpness of probabilistic and Euclidean errors, correlations, as well as correlations of errors of median/mean forecasts) the predictive performance of the bivariate EMOS model is compared to the forecast skills of the independent EMOS calibration of wind speed and temperature, the Gaussian copula method of Möller *et al.* (2013) based on univariate EMOS models, the bivariate BMA model suggested by Baran and Möller (2015) and the raw ensemble vectors as well.

From the results of the presented case studies one can conclude that compared to the raw ensemble post-processing always improves the calibration of probabilistic and accuracy of point forecasts. Further, in terms of predictive performance the bivariate EMOS model is able to keep up with the other two bivariate methods. Concerning the computational costs it outperforms the bivariate BMA method, whereas compared to the Gaussian copula approach it does not require an additional data set for estimating the correlations.

Acknowledgments. Essential part of this work was made during the visit of Sándor Baran at the Heidelberg Institute for Theoretical Studies. The research stay in Heidelberg was funded by the DAAD program "Research Stays for University Academics and Scientists, 2015". Sándor Baran was also supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The authors are indebted to Tilmann Gneiting for his useful suggestions and remarks, to the University of Washington MURI group for providing the UWME data, to Mihály Szűcs from the HMS for the ALADIN-HUNEPS data and to Thordis Thorarinsdottir and Alex Lenkoski for their help with the R codes for the copula method.

References

Bao, L., Gneiting, T., Raftery, A. E., Grimit, E. P. and Guttorp, P. (2010) Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon. Wea. Rev.* **138**, 1811–1821.

Baran, S. (2014) Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Comput. Stat. Data. Anal.* **75**, 227–238.

Baran, S., Horányi, A. and Nemoda, D. (2013) Statistical post-processing of probabilistic wind speed forecasting in Hungary. *Meteorol. Z.* **22**, 273–282.

Baran, S., Horányi, A. and Nemoda, D. (2014) Probabilistic temperature forecasting with statistical calibration in Hungary. *Meteorol. Atmos. Phys.* **124**, 129–142.

Baran, S., Lerch, S. (2015) Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2521.

Baran, S., Möller, A. (2015) Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics* **26**, 120–132.

Bouallègue, B. Z., Theis, S. and Gebhardt, C. (2013) Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorol. Z.* **22**, 49–59.

Buizza, R., Tribbia, J., Molteni, F. and Palmer, T. (1993) Computation of optimal unstable structures for a numerical weather prediction system. *Tellus A* **45**, 388–407.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.* **133**, 1076–1097.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X. and Stull, R. B. (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.* **111** D24307.

Dennis, J. and Schnabel, R. (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, New Jersey.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P. and Cébron, P. (2014). PEARP, the Météo-France short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2469.

Eckel, F. A. and Mass, C. F. (2005) Effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting* **20**, 328–350.

ECMWF Directorate (2012) Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter* **133**, 11–13.

Fraley, C., Raftery, A. E. and Gneiting, T. (2010) Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.* **138**, 190–202.

Fraley, C., Raftery, A. E., Gneiting, T., Sloughter, J. M. and Berrocal, V. J. (2011) Probabilistic weather forecasting in R. *The R Journal* **3**, 55–63.

Fritz, H., Filzmoser, P. and Croux, C. (2012) A comparison of algorithms for the multivariate L_1-median. *Comput. Stat.* **27**, 393–410.

Gneiting, T. (2014) Calibration of medium-range weather forecasts. *ECMWF Technical Memorandum* No. 719. Available at: old.ecmwf.int/publications/library/ecpublications/_pdf/tm/701-800/tm719.pdf. Accessed 27 July 2015.

Gneiting, T. and Raftery, A. E. (2005) Weather forecasting with ensemble methods. *Science* **310**, 248–249.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction and estimation. *J. Amer. Statist. Assoc.* **102**, 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.* **133**, 1098–1118.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008) Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test* **17**, 211–264.

Grell, G. A., Dudhia, J. and Stauffer, D. R. (1995) A description of the fifth-generation Penn state/NCAR mesoscale model (MM5). *Technical Note* NCAR/TN-398+STR. National Center for Atmospheric Research, Boulder. Available at: http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-214.pdf. Accessed 27 July 2015.

Horányi, A, Kertész, S., Kullmann, L. and Radnóti, G. (2006) The ARPEGE/ALADIN mesoscale numerical modelling system and its application at the Hungarian Meteorological Service. *Időjárás* **110**, 203–227.

Horányi, A., Mile, M., Szűcs, M. (2011) Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary. *Tellus A* **63**, 642–651.

Lee, G. and Scott, C. (2012) EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Comput. Statist. Data Anal.* **56**, 2816–2829.

Leith, C. E. (1974) Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev.* **102**, 409–418.

Lerch, S. and Thorarinsdottir, T. L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65**, 21206.

Leutbecher, M. and Palmer, T. N. (2008) Ensemble forecasting. *J. Comp. Phys.* **227**, 3515–3539.

Milasevic, P. and Ducharme, G. R. (1987) Uniqueness of the spatial median. *Ann. Statist.* **15**, 1332–1333.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.* **139**, 982–991.

National Weather Service (1998) *Automated Surface Observing System (ASOS) Users Guide.* Available at: http://www.weather.gov/asos/aum-toc.pdf. Accessed 27 July 2015.

Nelder, J. A. and Mead, R. (1965) A simplex algorithm for function minimization. *Comput. J.* **7**, 308–313.

Pinson, P. (2012) Adaptive calibration of (u, v)-wind ensemble forecasts. *Q. J. R. Meteorol. Soc.* **138**, 1273–1284.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. T. (2007) *Numerical Recipes 3rd Edition: The Art of Scientific Computing.* Cambridge University Press, Cambridge.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.* **133**, 1155–1174.

Rosenbaum, S. (1961) Moments of a truncated bivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **23**, 405–408.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28**, 616–640.

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* **149**, 1086–1096.

Schuhen, N., Thorarinsdottir, T. L. and Gneiting, T. (2012) Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.* **140**, 3204–3219.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.* **105**, 25–37.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2013) Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Mon. Wea. Rev.* **141**, 2107–2119.

Thorarinsdottir, T. L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* **173**, 371–388.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.* **125**, 3297–3319.

Vardi, Y. and Zhang, C. H. (2000) The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97**, 1423–1426.

Wilks, D. S. (2011) *Statistical Methods in the Atmospheric Sciences.* 3rd ed., Elsevier, Amsterdam.

Williams, R. M., Ferro, C. A. T. and Kwasniok, F. (2014) A comparison of ensemble post-processing methods for extreme events. *Q. J. R. Meteorol. Soc.* **140**, 1112–1120.